Population-Specific Genetic and Expression Differentiation in Europeans

Xueyuan Jiang[1] and Raquel Assis[1,2,3,4,*]

[1]Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802

[2]Department of Biology, Pennsylvania State University, University Park, PA 16802

[3]Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431

[4]Institute for Human Health and Disease Intervention, Florida Atlantic University, Boca Raton, FL 33431

*Corresponding author: E-mail: rassis@fau.edu

Accepted: January 29, 2020

Data deposition: This project has been deposited at (https://github.com/xueyuanj/human_PBS; last accessed February 12, 2020). In addition, the GitHub page contains all scripts used in the described analyses, as well as a README file explaining their usage.

Abstract

Much of the enormous phenotypic variation observed across human populations is thought to have arisen from events experienced as our ancestors peopled different regions of the world. However, little is known about the genes involved in these population-specific adaptations. Here, we explore this problem by simultaneously examining population-specific genetic and expression differentiation in four human populations. In particular, we derive a branch-based estimator of population-specific differentiation in four populations, and apply this statistic to single-nucleotide polymorphism and RNA-seq data from Italian, British, Finish, and Yoruban populations. As expected, genome-wide estimates of genetic and expression differentiation each independently recapitulate the known relationships among these four human populations, highlighting the utility of our statistic for identifying putative targets of population-specific adaptations. Moreover, genes with large copy number variations display elevated levels of population-specific genetic and expression differentiation, consistent with the hypothesis that gene duplication and deletion events are key reservoirs of adaptive variation. Further, many top-scoring genes are well-known targets of adaptation in Europeans, including those involved in lactase persistence and vitamin D absorption, and a handful of novel candidates represent promising avenues for future research. Together, these analyses reveal that our statistic can aid in uncovering genes involved in population-specific genetic and expression differentiation, and that such genes often play important roles in a diversity of adaptive and disease-related phenotypes in humans.

Key words: human evolution, population genetics, expression divergence, genetic divergence.

Introduction

Human phenotypes vary widely across the globe. In particular, geographically separated populations often differ in skin pigmentation (Loomis 1967), hair color (Rees 2003), tooth morphology (Scott and Turner 1997; Hanihara and Ishida 2005), surface area to body mass ratio (Katzmarzyk and Leonard 1998), and predisposition to diseases (Frank 2004). Much of this phenotypic variation is thought to have arisen due to a diversity of selective pressures experienced as early humans peopled the world and encountered novel environments (Sabeti et al. 2002; Voight et al. 2006), food sources (Sabeti et al. 2002), and pathogens (Diamond 2002; Jobling et al. 2013). As a result, uncovering the genetic targets of phenotypic differentiation among human populations is critical both for understanding past human adaptations (Sabeti et al. 2002) and for advancing future biomedical research (Jorde et al. 2001; Akey et al. 2004).

Due to the abundance of whole-genome sequence and polymorphism data for many human populations (Cann et al. 2002; International HapMap 3 Consortium 2010; 1000 Genomes Project Consortium 2015), much work in the past several years has focused on elucidating and understanding genetic differentiation that occurred during human evolution (Li et al. 2008; Pickrell et al. 2009; Field et al. 2016).

A common summary statistic for estimating genetic distances between two populations is the fixation index, F_{ST} (Wright 1951), which has been used to infer human demographic history (Hinds et al. 2005; Holsinger and Weir 2009; Keinan et al. 2009; Patterson et al. 2012; 1000 Genomes Project Consortium 2015) and to identify loci that may be targets of natural selection (Bowcock et al. 1991; Akey et al. 2002; Bersaglieri et al. 2004). However, because F_{ST} is a pairwise metric, it cannot identify the directionality of genetic differentiation nor be used as sole evidence for natural selection (Yi et al. 2010). To address this issue, Yi et al. (2010) developed the Population Branch Statistic (PBS), a summary statistic that utilizes pairwise F_{ST} values among three populations to quantify genetic differentiation along each branch of their corresponding three-population tree. Genes with large PBS values on one branch represent loci that underwent population-specific genetic differentiation consistent with relaxed selective constraint or positive selection (Yi et al. 2010). PBS has been applied to corroborate previously established targets of selection, including genes associated with skin pigmentation (Lamason et al. 2005) and dietary fat sources (Mathias et al. 2012), as well as to identify novel candidates for high-altitude adaptation in Tibetans (Yi et al. 2010).

However, because natural selection acts on phenotypes, analysis of genetic data only enables assessment of its indirect effects. For this reason, it may be advantageous to study selection more directly by exploiting the recent availability of RNA-seq data for several human populations (Lappalainen et al. 2013). Specifically, phenotypic evolution is thought to often occur through modifications in gene expression (King and Wilson 1975; Wang et al. 1996; Wray et al. 2003; Carroll 2005, 2008; Raj et al. 2010). Thus, studying gene expression differentiation among human populations may increase power for identifying loci underlying population-specific phenotypes. Indeed, like genetic differentiation, gene expression levels vary considerably across human populations (Cheung et al. 2005; Stranger et al. 2007) and often reflect population structure (Brown et al. 2018). Moreover, human genes with large PBS values are enriched for expression quantitative trait loci (Quiver and Lachance 2018).

In the present study, we simultaneously explore population-specific genetic and expression differentiation in four human populations: the Toscani in Italia (TSI), British in England and Scotland (GBR), Finnish in Finland (FIN), and Yoruba in Nigeria (YRI). For these analyses, we employ single-nucleotide polymorphism (SNP; 1000 Genomes Project Consortium 2015) and RNA-seq (Lappalainen et al. 2013) data from each population. First, we use F_{ST} (Wright 1951) and its analog for estimating quantitative trait differentiation, P_{ST} (Leinonen et al. 2006), to quantify and examine genome-wide patterns of genetic and expression differentiation in the four human populations. Next, we adapt the approach of PBS (Yi et al. 2010) to P_{ST}, and extend its computation to a four-population tree, enabling us to estimate both genetic and expression differentiation in each of the four human populations. Last, we apply this branch-based statistic to study population-specific genetic and expression differentiation, and uncover candidate genes and functional modules underlying adaptation in TSI, GBR, and FIN populations.

Results

Genome-Wide Patterns of Genetic and Expression Differentiation in Four Human Populations

A first goal of our study was to estimate genetic and expression differentiation among TSI, GBR, FIN, and YRI populations. To address this problem, we used SNP data (1000 Genomes Project Consortium 2015) to calculate the F_{ST} (Wright 1951), and RNA-seq data (Lappalainen et al. 2013) to calculate the P_{ST} (Leinonen et al. 2006), of every gene between each pair of the four human populations. We calculated F_{ST} using Hudson's formula (Hudson et al. 1992) and computed the ratio of averages to minimize bias (Reynolds et al. 1983; Weir and Cockerham 1984; International HapMap 3 Consortium 2010; Bhatia et al. 2013; see Materials and Methods for details). Due to environmental effects on P_{ST}, we followed the approach of Leinonen et al. (2006) in calculating P_{ST} under two contrasting scenarios: one in which environmental and nonadditive genetic effects account for half of the observed expression variation ($h^2 = 0.5$), and a second in which only additive genetic effects contribute to the observed expression variation ($h^2 = 1$; see Materials and Methods for details). Examinations of Pearson's linear (r) and Spearman's nonlinear (ρ) correlations revealed small ($\sim 10^{-2}$) but significantly positive relationships between F_{ST} and P_{ST} in TSI–FIN, TSI–YRI, GBR–YRI, and FIN–YRI population pairs (supplementary tables 1 and 2, Supplementary Material online), consistent with previous observations that genetic and expression differentiation are weakly or moderately associated (Makova and Li 2003; Nuzhdin et al. 2004; Sartor et al. 2006; Assis and Bachtrog 2013, 2015; Hunt et al. 2013).

To explore genome-wide patterns of genetic and expression differentiation among the four human populations, we independently used F_{ST} and P_{ST} to construct gene trees and then infer population trees supported by majorities of these gene trees (see Materials and Methods for details). Population trees inferred from F_{ST} and P_{ST} (with $h^2 = 0.5$ and $h^2 = 1$) have the same topology (fig. 1), indicating that there is consistency between relationships estimated from genome-wide patterns of genetic and expression differentiation despite their weak correlations with one another. Further, the topology of the inferred population trees recapitulates known relationships among these four populations, in that TSI and GBR are most closely related to one another, FIN is an outgroup to TSI and GBR, and YRI is an outgroup to all three European populations. These results mirror those from similar studies of F_{ST} (Hinds et al. 2005; Jakobsson et al. 2008; Li et al. 2008;

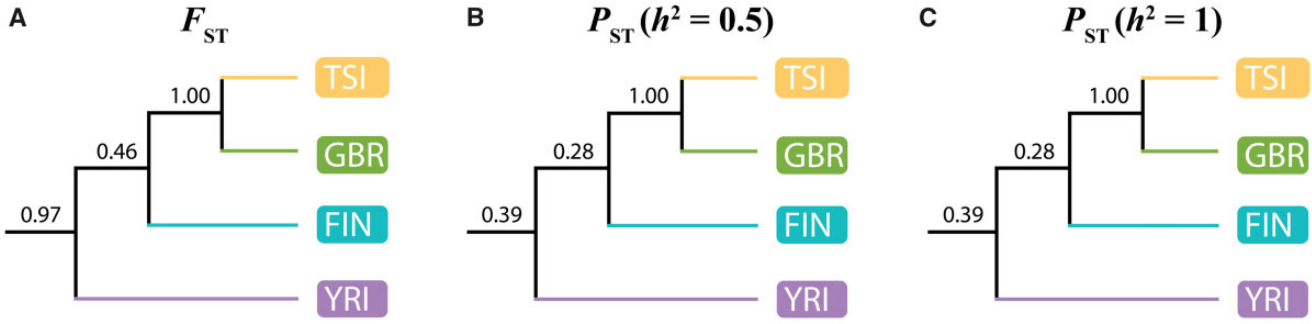

Fig. 1.—Relationships among TSI, GBR, FIN, and YRI populations inferred from genome-wide patterns of genetic and expression differentiation. Population trees supported by the majority of gene trees constructed using (*A*) F_{ST}, (*B*) P_{ST} with $h^2 = 0.5$, and (*C*) P_{ST} with $h^2 = 1$. Numbers indicate proportions of corresponding nodes in all gene trees (see Materials and Methods for details).

Auton et al. 2009; Holsinger and Weir 2009; Keinan et al. 2009; Patterson et al. 2012; 1000 Genomes Project Consortium 2015), as well as findings that gene expression data often display population structure comparable to that of genetic data (Cheung et al. 2005; Stranger et al. 2007; Brown et al. 2018).

Yet, there is greater support for the inferred population tree when using F_{ST} (fig. 1*A*) than when using P_{ST} (fig. 1*B* and *C*) as input. This effect is not surprising, given the complex and dynamic nature of gene expression data. Specifically, gene expression levels can vary across space (e.g., cell type), time (e.g., age), and condition (e.g., disease). Additionally, the experimental methodology used to collect and quantify these data may influence expression levels as well. This contrasts with the relatively static nature of genetic data. Further, whereas our calculation of F_{ST} for a gene was often based on allele frequencies at multiple SNPs across the gene, our calculation of P_{ST} for a gene was based on a single measurement. Therefore, differing levels of support observed for the inferred population trees may reflect higher accuracy and lower variance in estimating F_{ST} given the more representative and larger samples available for genetic data.

To investigate this effect, we examined the association between the number of SNPs in a gene and the difference between topologies of the gene tree constructed with F_{ST} and the population tree. In particular, if mismatches between gene trees constructed with P_{ST} and the population tree are often due to the small sample size of expression data, then we also expect gene trees constructed with F_{ST} to be different from the population tree when the number of SNPs is small. To quantify the difference between each gene tree constructed with F_{ST} and the population tree, we used the Robinson–Foulds (RF) distance, which is the sum of the number of unique clades in the two trees being compared (Robinson and Foulds 1981). Here, RF $= 0$ when the tree topologies are identical, RF $= 2$ when there is one unique clade in each tree, and RF $= 4$ when the tree topologies are distinct. As hypothesized, there is an inverse relationship between RF and the number of SNPs, in that we tend to get RF $= 0$ when the number of SNPs is largest, RF $= 2$ when the number of SNPs is intermediate, and RF $= 4$ when the number of SNPs is smallest (supplementary fig. 1, Supplementary Material online; $P < 0.01$ for all pairwise comparisons, two-sample permutation tests; see Materials and Methods for details). Hence, whereas genome-wide patterns of genetic and expression differentiation likely reflect true population relationships (fig. 1), gene-level values of F_{ST}, and particularly of P_{ST}, should be interpreted with caution.

Estimation of Population-Specific Genetic and Expression Differentiation on a Four-Population Tree

Next, we sought to quantify population-specific genetic and expression differentiation of genes in each of the four human populations. For a three-population tree, population-specific genetic differentiation of a gene along each branch can be estimated with PBS (Yi et al. 2010; fig. 2*A*), which applies equation (11.20) in Felsenstein (2004) to F_{ST}. In particular, considering the unrooted three-population tree shown in figure 2*A*, the PBS value of a particular gene in population W is estimated as $\text{PBS}_W = \frac{1}{2}(E_{W,X} + E_{W,Y} - E_{X,Y})$, where $E_{W,X}$, $E_{W,Y}$, and $E_{X,Y}$ denote the log-transformed F_{ST} between populations W and X, W and Y, and X and Y, respectively (Yi et al. 2010; see Materials and Methods for details). In a recent study, equation (11.20) in Felsenstein (2004) was also applied to expression distances between orthologous genes to estimate branch lengths corresponding to lineage-specific expression divergence on a three-species tree (Assis 2019). Analogously, by substituting P_{ST} for F_{ST} in the formula for PBS (Yi et al. 2010), we can obtain the PBS corresponding to gene expression differentiation in population W on the three-population tree. To distinguish between these two PBS in our study, we will refer to the calculation with F_{ST} as "genetic PBS," and the calculation with P_{ST} as "expression PBS."

To enable quantification of population-specific genetic and expression differentiation in four human populations, we

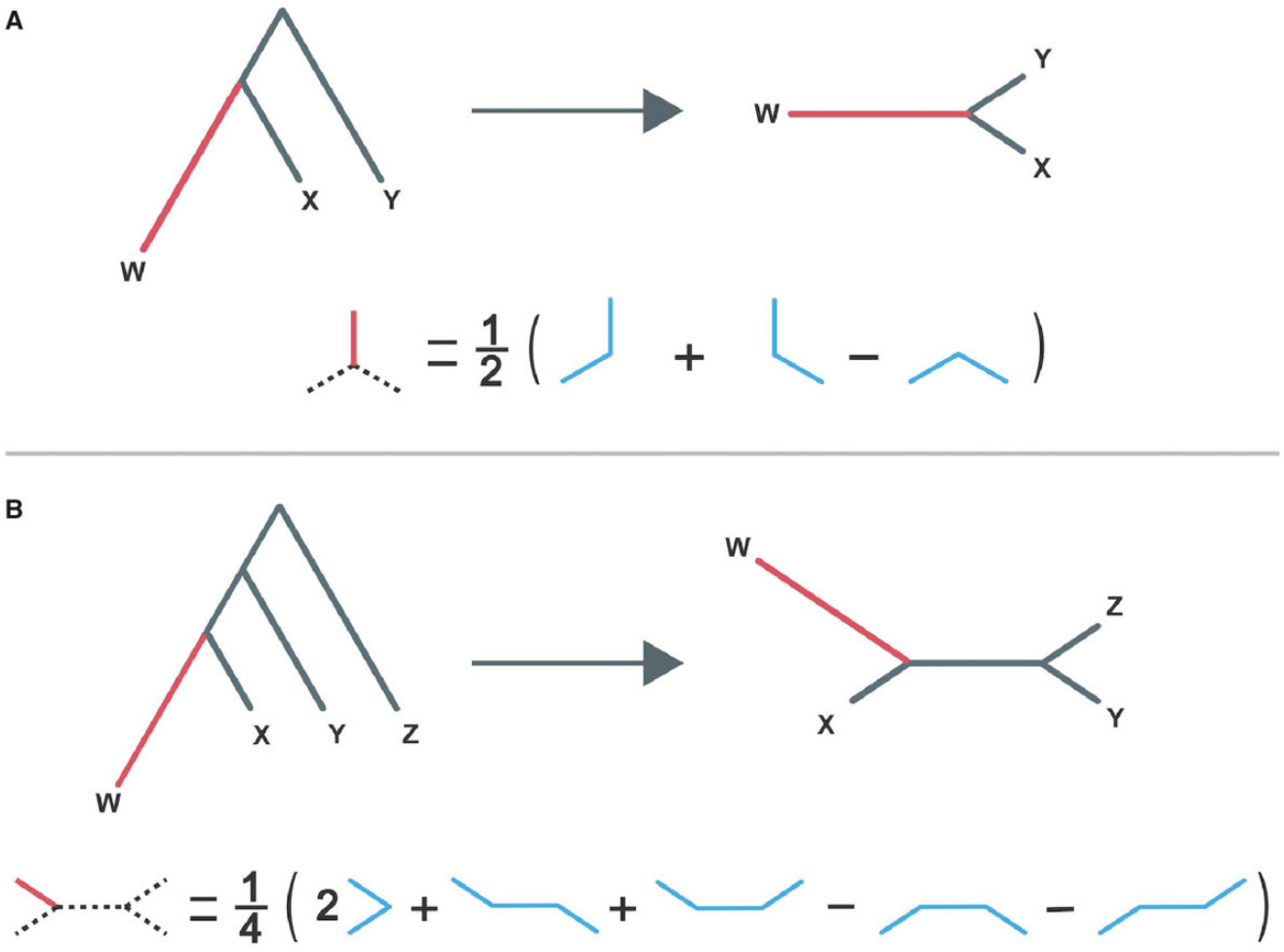

Fig. 2.—Schematic for calculating the PBS value of a gene in population *W*. Depicted are scenarios in which population-specific differentiation of a gene has occurred in population *W* of a set of (*A*) three populations *W*, *X*, and *Y* and (*B*) four populations *W*, *X*, *Y*, and *Z*. In each case, population-specific differentiation results in elongation of external branch *W* (red). To estimate the length of external branch *W*, we unroot the tree (top of each panel) and apply the formula shown (bottom of each panel) to pairwise genetic (F_{ST}) or expression (P_{ST}) distances between populations. We can use an analogous approach to estimate lengths of other external branches.

extended the derivation of PBS to a four-population tree (fig. 2*B*). Henceforth, we will denote PBS as PBS_3 when applied to a three-population tree (fig. 2*A*) and as PBS_4 when applied to a four-population tree (fig. 2*B*). To derive PBS_4, suppose that we have four populations *W*, *X*, *Y*, and *Z* that are related by the unrooted tree depicted in figure 2*B*. Then, we can compute four PBS_4 values for a particular gene, one corresponding to its population-specific differentiation in each population. Because the PBS_4 value for a gene in a population represents its differentiation that occurred in the lineage of that population, it can be estimated by the length of the external branch corresponding to the population. We can obtain the length of each external branch by first computing four distances: those between populations *W* and *X* ($E_{W,X}$), *W* and *Y* ($E_{W,Y}$), *X* and *Y* ($E_{X,Y}$), and *X* and *Z* ($E_{X,Z}$). Then, we can use these distances to compute the length of each external branch by following the schematic pictured in figure 2*B*. For example, the PBS_4 value of the gene in population *W* is calculated as $PBS_{4,W} = \frac{1}{4}(2E_{W,X} + E_{W,Y} + E_{W,Z} - E_{X,Y} - E_{X,Z})$. Using this formula, we computed the genetic PBS_4 and expression PBS_4 of each gene in TSI, GBR, FIN, and YRI populations (supplementary tables 3–5, Supplementary Material online; see Materials and Methods for details).

Population-Specific Genetic and Expression Differentiation of Genes with Copy Number Variations

Gene duplications and deletions are key contributors to human genetic diversity (Sudmant et al. 2015). Moreover, because they are large-scale mutation events that may impact gene dosage, duplications and deletions have been implicated in numerous human diseases (Sebat et al. 2004; Kumar et al. 2008; Sharp et al. 2008; Weiss et al. 2008), as well as in adaptive events in many diverse species (Kaessmann 2010;

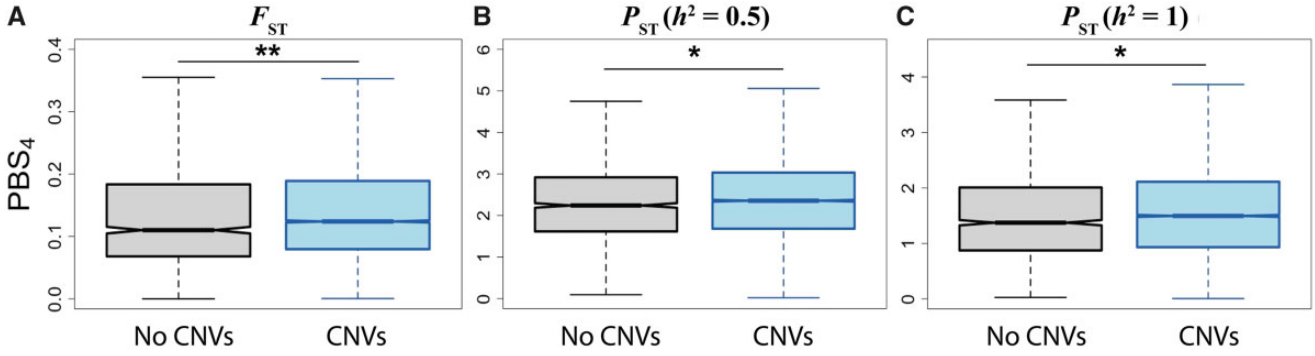

Fig. 3.—PBS_4 values of genes with CNVs. Distributions of (*A*) genetic PBS_4 values calculated from F_{ST}, (*B*) expression PBS_4 values calculated from P_{ST} with $h^2 = 0.5$, and (*C*) expression PBS_4 values calculated from P_{ST} with $h^2 = 1$ of genes without (gray) and with (blue) CNVs. $*P < 0.05$ and $**P < 0.001$ (see Materials and Methods for details).

Chen et al. 2013). For these reasons, genes harboring copy number variations (CNVs) are thought to be more frequently targeted by natural selection than those without CNVs (Freeman et al. 2006; Nguyen et al. 2006). Indeed, genes with CNVs often display signatures of adaptation (Sudmant et al. 2015), and fixation of duplications and deletions has been associated with natural selection in many species (Freeman et al. 2006; Nguyen et al. 2006; Han, Demuth, et al. 2009; Jiang and Assis 2017). Therefore, we hypothesized that genes with CNVs would have larger genetic and expression PBS_4 values than genes without CNVs. To test this hypothesis, we compared the distributions of maximum PBS_4 values of genes with and without known human CNVs larger than 50 bp (fig. 3; MacDonald et al. 2014; see Materials and Methods for details). As expected, both genetic and expression PBS_4 values are significantly elevated in genes with CNVs (fig. 3; $P < 0.05$ for all pairwise comparisons, two-sample permutation tests; see Materials and Methods for details). Though the magnitudes of the effects are modest, genes with CNVs also contain more SNPs than those without CNVs ($P < 0.001$, two-sample permutation test; see Materials and Methods for details), which is expected to decrease their genetic PBS_4 values (Yi et al. 2010). Taken together, these findings suggest that genes with CNVs tend to undergo increased population-specific genetic and expression differentiation that is consistent with positive selection.

However, increased population-specific genetic and expression differentiation of genes with CNVs may not only be attributed to positive selection, but alternatively to relaxed selective constraint. To disentangle these mechanisms, we examined levels of background selection in genes with and without CNVs. Background selection reduces genetic diversity at linked deleterious sites (Charlesworth et al. 1993), and is therefore weaker in regions with reduced selective constraint. As a result, if genes with CNVs primarily evolve under relaxed selective constraint, then we expect a reduction in their levels of background selection relative to those of genes without CNVs. To determine whether this is the case, we compared distributions of median *B* values (McVicker et al. 2009) in genes with and without CNVs. We found no significant difference between groups (supplementary fig. 2*A*, Supplementary Material online, $P > 0.05$, two-sample permutation test; see Materials and Methods for details), suggesting that overall levels of selective constraint do not differ between genes with and without CNVs. Further, because F_{ST} is correlated with background selection (Charlesworth et al. 1997), we performed a follow-up analysis in which we explicitly accounted for background selection when comparing the genetic PBS_4 of genes with and without CNVs. Specifically, we corrected F_{ST} for background selection using estimated *B* values (see supplementary Methods, Supplementary Material online, for derivation) and recalculated the background selection-corrected F_{ST} and genetic PBS_4 of each gene. Even after this correction, genetic PBS_4 is elevated in genes with CNVs (supplementary fig. 2*B*, Supplementary Material online, $P < 0.001$, two-sample permutation test; see Materials and Methods for details). Whereas *B* values are not perfect measures of selective constraint, particularly for short evolutionary timescales, these findings better support the hypothesis that increased population-specific differentiation in genes with CNVs is due to positive selection than to relaxed selective constraint.

Relationship of Population-Specific Genetic and Expression Differentiation to Gene Function in Europeans

A natural question that emerges from our study is whether there are functional drivers of population-specific genetic and expression differentiation. In answering this question, it was important to exclude YRI, as it is an outgroup to the three European populations and therefore contains greater overall population-specific genetic and expression differentiation that cannot be polarized. Hence, we only considered TSI, GBR, and FIN populations. To globally assess functional modules contributing to population-specific genetic and expression differentiation in these populations, we utilized annotation data from the GO Consortium (Ashburner et al. 2000; GO Consortium

2018). In particular, GO terms classify genes by their molecular functions, cellular components, and biological processes (Ashburner et al. 2000; GO Consortium 2018). Though GO terms refer to intracellular gene functions that cannot be directly related to phenotypes that natural selection acts on, they can aid in elucidating the classes of gene functions that may be associated with population-specific genetic and expression differentiation. To examine these associations, we ranked genes by their genetic and expression PBS_4 values in each European population, performed GO enrichment analysis on ranked lists, and extracted significantly overrepresented GO terms (supplementary tables S6–S14, Supplementary Material online; see Materials and Methods for details).

After correcting for multiple testing, there are no significantly enriched GO terms for genetic PBS_4 in any of the populations (supplementary tables S6–S8, Supplementary Material online). However, there are many significantly enriched GO terms for expression PBS_4 in all three populations (supplementary tables S9–S14, Supplementary Material online). Enriched GO terms for expression PBS_4 calculated from P_{ST} with $h^2 = 0.5$ and $h^2 = 1$ are similar, consistent with our previous comparisons (see figs. 1 and 3). Moreover, several enriched GO terms are shared among the three related populations, and numerous related terms are enriched in individual populations. Though most GO terms are quite general and have limited interpretability, it appears that population-specific expression differentiation in Europeans often affects genes involved in signal transduction and immunity. This is not surprising, as such processes are frequent targets of natural selection (Barreiro and Quintana-Murci 2010; Fumagalli et al. 2011; Enard et al. 2016).

To glean further insight into the individual genes potentially driving population-specific genetic and expression differentiation in Europeans, we performed literature searches on genes with the largest genetic and expression PBS_4 values in each population (tables 1 and 2). In both TSI and GBR, the gene with the largest genetic PBS_4 value is *MCM6*, or Minichromosome Maintenance Complex Component 6. *MCM6* is part of a protein complex essential for the initiation of eukaryotic genome replication (Labib et al. 2000). Two of its introns contain enhancers for its upstream gene *LCT*, or Lactase, one of which has a mutation prevalent in European populations that is thought to confer lactose tolerance in adulthood (Enattah et al. 2002; Troelsen et al. 2003). Interestingly, *LCT* also has the second-largest genetic PBS_4 in GBR, and several genetic studies have identified both *MCM6* and *LCT* as targets of recent positive selection in Europeans (Bersaglieri et al. 2004; Voight et al. 2006; Ranciaro et al. 2014; Cheng et al. 2017). In FIN, the gene with the largest genetic PBS_4 value is *HLA-DPA1*, or Major Histocompatibility Complex, Class II, DP Alpha 1. As a member of the *HLA* gene family, *HLA-DPA1* plays an important role in antigen presentation (Bottazzo et al. 1983) and is believed to be evolving under balancing selection in humans (Hughes and Nei 1988, 1989; Takahata and Nei 1990; Hughes and Yeager 1998; Yasukochi and Satta 2013).

Table 1
Genes with Top Five Genetic PBS_4 Values in TSI, GBR, and FIN

	TSI	GBR	FIN
1	*MCM6*	*MCM6*	*HLA-DPA1*
2	*DCUN1D4*	*LCT*	*RNF114*
3	*DARS*	*CCNT2*	*TRIM47*
4	*CCNT2*	*R3HDM1*	*HSPA2*
5	*PRDM4*	*ZNF615*	*FAHD2B*

Table 2
Genes with Top Five Expression PBS_4 Values (P_{ST} with $h^2 = 0.5$ and $h^2 = 1$) in TSI, GBR, and FIN

	TSI		GBR		FIN	
	$h^2 = 0.5$	$h^2 = 1$	$h^2 = 0.5$	$h^2 = 1$	$h^2 = 0.5$	$h^2 = 1$
1	*PRKCB*	*PRKCB*	*PRRX1*	*PRRX1*	*VDR*	*FZD1*
2	*TBC1D1*	*TBC1D1*	*CD28*	*CD28*	*FZD1*	*VDR*
3	*BMPR1A*	*KLF3*	*MOB1B*	*INSR*	*TMEM144*	*PLAC8*
4	*KLF3*	*MGAT5*	*BTBD3*	*BTBD3*	*ACTN1*	*FAM134B*
5	*MGAT5*	*FAM65B*	*GLDC*	*TBXT*	*PLAC8*	*SYNJ2*

In TSI, the gene with the largest expression PBS_4 value (calculated from P_{ST} with $h^2 = 0.5$ and $h^2 = 1$) is *PRKCB*, or Protein Kinase C Beta. *PRKCB* is involved in numerous signaling pathways, including apoptosis (Reyland 2009) and B cell activation during immune response (Lutzny et al. 2013). As a result, mutations in *PRKCB* are associated with many cancers (Lutzny et al. 2013; Wallace et al. 2014; Antal et al. 2015) and autoimmune diseases (Han, Zheng, et al. 2009; Sheng et al. 2011; Kawashima et al. 2017). The association with autoimmune diseases is particularly intriguing, as such genes are often targets of recent positive selection (Barreiro and Quintana-Murci 2010; Ramos et al. 2014). It is hypothesized that mutations that cause autoimmune response today may have conferred pathogen resistance in the past (Barreiro and Quintana-Murci 2010). In GBR, the gene with the largest expression PBS_4 value (calculated from P_{ST} with $h^2 = 0.5$ and $h^2 = 1$) is *PRRX1*, or Paired Related Homeobox 1. *PRRX1* is a DNA-associated protein that is involved in the establishment of diverse mesodermal muscle types during development (Martin et al. 1995). It has also been connected with numerous cancers (Takahashi et al. 2013; Guo et al. 2015; Hirata et al. 2015; Jurecekova et al. 2016; Takano et al. 2016; Zhu et al. 2017) and is thought to mediate metastasis (Ocaña et al. 2012; Takahashi et al. 2013; Guo et al. 2015; Zhu et al. 2017). In FIN, the genes with the two largest expression PBS_4 values are *VDR* followed by *FZD1* when P_{ST} was calculated with $h^2 = 0.5$, and *FZD1* followed by *VDR* when P_{ST} was calculated with $h^2 = 1$. *VDR*, or Vitamin D Receptor, interacts with vitamin D in the small intestine to facilitate calcium transportation into

circulation (Holick 2006). Skin exposure to solar ultraviolet radiation produces about 90% of the vitamin D that an individual requires (Holick 2006), and living at high latitudes has been associated with vitamin D deficiency due to decreased ultraviolet radiation (Kimlin 2008; Chaplin and Jablonski 2009). Therefore, it is possible that expression differentiation of *VDR* may contribute to high latitude adaptation in FIN. *FZD1*, or Frizzled Class Receptor 1, is a receptor for Wnt signaling proteins (Kennerdell and Carthew 1998). It has been associated with several cancers (Kirikoshi et al. 2001; Benhaj et al. 2006; Zhang et al. 2015) and specifically with chemoresistance (Flahaut et al. 2009), thus making it a promising therapeutic target.

Materials and Methods

Gene Expression Analyses

We obtained RNA-seq data from lymphoblastoid cell lines in TSI, GBR, FIN, and YRI populations from the GEUVADIS project (Lappalainen et al. 2013). These data comprise 93 individuals in TSI, 94 individuals in GBR, 95 individuals in FIN, and 89 individuals in YRI, all of whom are from the 1000 Genomes Project (1000 Genomes Project Consortium 2015). We excluded data from the population of Utah Residents with Northern and Western European Ancestry (CEU) because they were collected from an older cell line and have been shown to display expression patterns that are inconsistent with their relationships to other populations (Yuan et al. 2015). We quantified the abundance of transcripts using featureCounts (Liao et al. 2014) with default parameters and the GRCh37 human genome (Zerbino et al. 2018) as our reference. To normalize count data, we used the "median ratio" method (Anders and Huber 2010) by implementing the estimateSizeFactors function in DESeq2 (Love et al. 2014). Next, we calculated the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) of each gene using DESeq2 (Love et al. 2014). We removed genes that contained fewer than ten reads in each sample (lowly expressed), were located on sex chromosomes, or were not protein coding. For the remaining 13,075 genes, we log-transformed their FPKM values by log(FPKM + 1). We computed the P_{ST} for each gene as $P_{ST} = \frac{\sigma^2_{between}}{\sigma^2_{between} + 2h^2\ \sigma^2_{within}}$ (Leinonen et al. 2006), where $\sigma^2_{between}$ is expression variance between populations, σ^2_{within} is expression variance within populations, and h^2 is heritability. For our analysis, we used $h^2 = 0.5$ and $h^2 = 1$ as was done previously (Leinonen et al. 2006), though we note that the patterns in figure 1 do not change as a function of h^2. When $h^2 = 1$, P_{ST} reduces to Q_{ST} (Spitze 1993), another common metric for differentiation of quantitative traits between populations.

Population-Genetic Analyses

We downloaded the 1000 Genomes Project phase 3 data set (1000 Genomes Project Consortium 2015) for TSI, GBR, FIN, and YRI populations from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/, last accessed February 12, 2020. To be conservative in our analyses, we only included the 371 individuals also present in the GEUVADIS Project (Lappalainen et al. 2013). After filtering out insertions, deletions, and monomorphic sites, we were left with 30,734,317 biallelic SNPs. Though we used SNPs of all allele frequencies, limiting our analysis to those with minor allele frequencies >0.01 did not alter our findings. We calculated Hudson's F_{ST} for each SNP as $F_{ST}^{Hudson} = \frac{(p_1-p_2)^2 - \frac{p_1(1-p_1)}{n_1-1} - \frac{p_2(1-p_2)}{n_2-1}}{p_1(1-p_2)+p_2(1-p_1)}$ (Reynolds et al. 1983; Weir and Cockerham 1984; Bhatia et al. 2013). Then, we combined SNPs within the entire annotated region of each gene and computed the "ratio of averages" for Hudson's F_{ST} (Reynolds et al. 1983; Weir and Cockerham 1984; Bhatia et al. 2013). Because negative F_{ST} values are not defined (Wright 1951) and have no biological interpretation (Akey et al. 2002), we followed the standard of setting all negative $F_{ST} = 0$ (e.g., Nei 1990; Akey et al. 2002).

Phylogenetic Analyses

To infer population trees, we first constructed gene trees using the NEIGHBOR program in the PHYLIP package (Felsenstein 2005). We constructed gene trees using either F_{ST} or P_{ST} as input distances between populations. Application of the UPGMA algorithm in the NEIGHBOR program yielded totals of 12,977 gene trees for F_{ST} and 13,075 gene trees for P_{ST}. Next, we used gene trees as input for the CONSENSE program in the PHYLIP package (Felsenstein 1993) and obtained rooted population trees supported by the majority of gene trees based on F_{ST} and P_{ST}. Specifically, the nodes in gene trees are included if they continue to resolve the population tree and do not contradict with more frequently occurring nodes. The number above each node in figure 1 represents its proportion in all gene trees.

Calculation of PBS_4

We first computed the genetic or expression distance between populations as $E_{A,B} = -\log[1 - Z_{ST}(A, B)]$, following the approach of Cavalli-Sforza (1969), where Z_{ST} represents either F_{ST} or P_{ST} between populations A and B. We used these as input for calculations of genetic and expression PBS_4 values. Negative branch lengths were set to 0.

Gene Ontology Enrichment Analyses

Genes were ranked by their genetic PBS_4 and expression PBS_4 values in each population (provided in supplementary tables S3–S5, Supplementary Material online). We performed Gene Ontology (GO) enrichment analysis on each ranked list of genes with the web-based GOrilla tool at http://cbl-gorilla.cs.technion.ac.il/; last accessed February 12, 2020 (Eden et al. 2007, 2009), which searches for enriched GO terms

that appear densely at the top of a ranked list of genes (Eden et al. 2007, 2009). For each run, we chose "*Homo sapiens*" as the organism, set the running mode to "Single ranked list of genes," selected all ontologies (process, function, and component), and set the threshold $P = 10^{-3}$.

Statistical Analyses

All statistical analyses were performed in the R software environment (R Core Team 2013). Two-sample permutation tests were used to assess differences between all pairs of distributions compared in figure 3 and supplementary figures 1 and 2, Supplementary Material online. For each test, we performed 1,000 permutations, using the difference between medians of groups as the test statistic. In particular, we computed the difference between the medians of the two groups for each permutation, and the *P* value of the permutation test as the proportion of times the absolute value of this difference was greater than or equal to the absolute value of the observed difference in the data. Student's *t*-tests were used to assess the statistical significance of correlation coefficients shown in supplementary tables 1 and 2, Supplementary Material online.

Discussion

Identifying drivers of human phenotypic differentiation is crucial to understanding adaptive events that occurred in the past, as well as to developing population- and individual-targeted treatments for diseases in the future (Jorde et al. 2001; Sabeti et al. 2002; Akey et al. 2004). Though previous research (Sabeti et al. 2002; Akey et al. 2004; Voight et al. 2006) has made use of abundant whole-genome and polymorphism data for many human populations (International HapMap 3 Consortium 2010; 1000 Genomes Project Consortium 2015) to answer this question, simultaneously studying genetic and expression differentiation may provide unique insights into direct phenotypic targets of natural selection. In particular, it is thought that phenotypic evolution more often occurs through changes in gene regulation and expression, rather than their protein-coding sequences (King and Wilson 1975; Wang et al. 1996; Wray et al. 2003; Carroll 2005, 2008; Raj et al. 2010). For this reason, gene expression differentiation might better reflect phenotypic differentiation. Therefore, a major advantage of the present study is that we utilized both genetic and expression data to address questions about population-specific differentiation in humans. Further, results from our combined analysis suggest that population-specific genetic and expression differentiation in humans may be attributed to several important biological processes, most notably signal transduction and immunity, and also pinpoint many candidate genes for future studies of human phenotypic variation in adaptation and disease.

Yet, there are three key limitations of the data analyzed here that must be considered when interpreting our findings in the context of human evolution. The first is that there is only a single estimate of the expression level of a gene in each population, which is particularly problematic given the complex and dynamic nature of gene expression data. In contrast, there are multiple SNPs per gene in each population, and genetic data are static. Therefore, we expect our estimates derived from expression data to have lower accuracy and higher variance than those from genetic data. Indeed, we found that gene trees constructed with F_{ST} match the topology of the inferred population tree more often than those constructed with P_{ST} and, further, that mismatches between topologies of gene trees constructed with F_{ST} and the inferred population tree are associated with fewer SNPs. Hence, it is also not surprising that genetic and expression PBS_4 do not have common outlier genes (supplementary tables S3–S5, Supplementary Material online), and gene-level values of expression (and in some cases genetic) PBS_4 should thus be interpreted with caution. In spite of this issue, a handful of genes with the largest expression PBS_4 are well-known candidates of adaptation, such as *VDR* (Kimlin 2008; Chaplin and Jablonski 2009). Moreover, at a genome-wide level, the discordance between findings derived from genetic and expression data illustrates the importance of integrating both types of data into population-genetic studies. Nevertheless, future availability of larger sample sizes for gene expression data in multiple human populations will be invaluable for accurately pinpointing genic targets of population-specific expression differentiation in humans.

The second caveat is that TSI, GBR, and FIN are closely related European populations. As a result, genetic distances among them are small, which can lead to noise in gene-level analyses. Moreover, due to shared ancestry and gene flow among these closely related populations, their genetic and expression differentiation are likely to be correlated. This limitation is clearly demonstrated by *MCM6* having the largest genetic PBS_4 value in both TSI and GBR, which are the most closely related of the three European populations studied. Thus, though genome-wide patterns of genetic and expression differentiation are consistent with population relationships, caution needs to be taken when making inferences based on the genetic and expression PBS_4 values of individual genes. Despite this limitation, several genes with the largest genetic PBS_4 values, such as *MCM6* and *HLA-DPA1*, are well-established targets of natural selection (Hughes and Nei 1988, 1989; Takahata and Nei 1990; Hughes and Yeager 1998; Bersaglieri et al. 2004; Voight et al. 2006; Yasukochi and Satta 2013; Ranciaro et al. 2014; Cheng et al. 2017), and novel candidates therefore may represent promising avenues for future research. Nevertheless, phenotypic differences among distantly related populations are better described than those among closely related populations, making it inherently more difficult to interpret our findings in the context of human phenotypes. Therefore, future availability of RNA-seq data from

additional populations, particularly those that are more distantly related, will be critical to studying population-specific variation and its role in both human evolution and disease.

The third limitation is that the RNA-seq data used in this study were obtained from lymphoblastoid cell lines. In particular, the enrichment of immune-related functions in genes with high levels of population-specific expression differentiation may be attributed to usage of this cell line, rather than reflecting widespread evolutionary patterns of immunity genes across tissues. Yet, it is important to note that associations between increased population-specific expression differentiation and immunity are consistent with previous findings. Specifically, immunity genes are among the fastest evolving genes in the human genome, likely due to adaptations to rapidly changing environments and introductions of novel pathogens (Barreiro and Quintana-Murci 2010; Fumagalli et al. 2011; Enard et al. 2016). Therefore, though observed patterns of population-specific expression differentiation may not be representative of those in other cell types, genes with high population-specific expression differentiation should be further studied to examine their potential roles in human evolutionary history and disease. Regardless, future availability of RNA-seq data for multiple cell or tissue types in several populations will be invaluable for capturing complex patterns of population-specific expression differentiation and pinpointing genic targets of phenotypic variation among human populations.

In spite of the noted issues with the data analyzed here, a major advantage of our study is the design of PBS_4, a novel summary statistic that can be used to estimate population-specific differentiation of a quantitative trait in four populations. PBS_4 requires minimal assumptions about the data and can be used to rapidly estimate population-specific differentiation on a genome-wide scale. Further, because PBS_4 utilizes data from four populations, branch lengths are more likely to represent true population-specific differentiation than differentiation that occurred ancestral to two populations, as is possible in a three-population scenario (Assis 2019). Therefore, though the data set used in our study is not ideal in many respects, PBS_4 can easily be applied to existing or future data sets to estimate population-specific differentiation of a wide array of genetic, expression, and other measurable traits in humans and other species. In particular, we envision that application of PBS_4 to future human RNA-seq data from multiple cell lines or tissues and in many populations of varying divergence levels will shed light on complex questions about human evolutionary history and disease processes.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the National Science Foundation (DEB-1555981). Portions of this research were conducted with Advanced Cyber Infrastructure computational resources provided by the Institute for CyberScience at Pennsylvania State University (https://ics.psu.edu; last accessed February 12, 2020). We would also like to thank the Associate Editor and three anonymous reviewers for their helpful comments during the review process.

Literature Cited

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526:68–74.
Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 12(12):1805–1814.
Akey JM, et al. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. 2(10):e286.
Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol. 11(10):R106.
Antal CE, et al. 2015. Cancer-associated protein kinase C mutations reveal kinase's role as tumor suppressor. Cell 160(3):489–502.
Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. Nat Genet. 25(1):25–29.
Assis R. 2019. Lineage-specific expression divergence in grasses is associated with male reproduction, host-pathogen defense, and domestication. Genome Biol Evol. 11(1):207–219.
Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in *Drosophila*. Proc Natl Acad Sci U S A. 110(43):17409–17414.
Assis R, Bachtrog D. 2015. Rapid divergence and diversification of mammalian duplicate gene functions. BMC Evol Biol. 15(1):138.
Auton A, et al. 2009. Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res. 19(5):795–803.
Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet. 11(1):17–30.
Benhaj K, Akcali KC, Ozturk M. 2006. Redundant expression of canonical Wnt ligands in human breast cancer cell lines. Oncol Rep. 15:701–707.
Bersaglieri T, et al. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet. 74(6):1111–1120.
Bhatia G, Patterson NJ, Sankararaman S, Price AL. 2013. Estimating and interpreting F_{ST}: the impact of rare variants. Genome Res. 23(9):1514–1521.
Bottazzo G, Hanafusa T, Pujol-Borrell R, Feldmann M. 1983. Role of aberrant HLA-DR expression and antigen presentation in induction of endocrine autoimmunity. Lancet 322(8359):1115–1119.
Bowcock AM, et al. 1991. Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc Natl Acad Sci U S A. 88(3):839–843.
Brown BC, Bray NL, Pachter L. 2018. Expression reflects population structure. PLoS Genet. 14(12):e1007841.
Cann HM, et al. 2002. A human genome diversity cell line panel. Science 296(5566):261–262.
Carroll SB. 2005. Evolution at two levels: on genes and form. PLoS Biol. 3(7):e245.
Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134(1):25–36.
Cavalli-Sforza LL. 1969. Human diversity. In: Proceedings of the 12th International Congress on Genetics. Vol. 2. p. 405–416.

Chaplin G, Jablonski NG. 2009. Vitamin D and the evolution of human depigmentation. Am J Phys Anthropol. 139(4):451–461.
Charlesworth B, Morgan M T, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics. 134(4):1289–1303.
Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res. 70(2):155–174.
Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. Nat Rev Genet. 14(9):645–660.
Cheng X, Xu C, DeGiorgio M. 2017. Fast and robust detection of ancestral selective sweeps. Mol Ecol. 26(24):6871–6891.
Cheung VG, et al. 2005. Mapping determinants of human gene expression by regional and genome-wide association. Nature 437(7063):1365–1369.
Diamond J. 2002. Evolution, consequences and future of plant and animal domestication. Nature 418(6898):700–707.
Eden E, Lipson D, Yogev S, Yakhini Z. 2007. Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol. 3(3):e39.
Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10(1):48.
Enard D, Cai L, Gwennap C, Petrov DA. 2016. Viruses are a dominant driver of protein adaptation in mammals. Elife 5:e12469.
Enattah NS, et al. 2002. Identification of a variant associated with adult-type hypolactasia. Nat Genet. 30(2):233–237.
Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, Seattle: University of Washington.
Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.
Field Y, et al. 2016. Detection of human adaptation during the past 2000 years. Science 354(6313):760–764.
Flahaut M, et al. 2009. The Wnt receptor FZD1 mediates chemoresistance in neuroblastoma through activation of the Wnt/β-catenin pathway. Oncogene 28(23):2245–2256.
Frank SA. 2004. Genetic predisposition to cancer—insights from population genetics. Nat Rev Genet. 5(10):764–772.
Freeman JL, et al. 2006. Copy number variation: new insights in genome diversity. Genome Res. 16(8):949–961.
Fumagalli M, et al. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. 7(11):e1002355.
GO Consortium. 2018. The Gene Ontology resource: 20 years and still GOing strong. Nucleic Acids Res. 47:D330–D338.
Guo J, et al. 2015. PRRX1 promotes epithelial–mesenchymal transition through the Wnt/β-catenin pathway in gastric cancer. Med Oncol. 32(1):393.
Han J-W, Zheng H-F, et al. 2009. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat Genet. 41(11):1234–1237.
Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW. 2009. Adaptive evolution of young gene duplicates in mammals. Genome Res. 19(5):859–867.
Hanihara T, Ishida H. 2005. Metric dental variation of major human populations. Am J Phys Anthropol. 128(2):287–298.
Hernandez RD, et al. 2011. Classic selective sweeps were rare in recent human evolution. Science 331(6019):920–924.
Hinds DA, et al. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307(5712):1072–1079.
Hirata H, et al. 2015. Downregulation of PRRX1 confers cancer stem cell-like properties and predicts poor prognosis in hepatocellular carcinoma. Ann Surg Oncol. 22(S3):1402–1409.
Holick MF. 2006. Resurrection of vitamin D deficiency and rickets. J Clin Invest. 116(8):2062–2072.
Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nat Rev Genet. 10(9):639–650.
Hudson RR, Slatkin M, Maddison W. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132(2):583–589.
Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335(6186):167–170.
Hughes AL, Nei M. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Nat Acad Sci U S A. 86(3):958–962.
Hughes AL, Yeager M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. Annu Rev Genet. 32(1):415–435.
Hunt BG, Ometto L, Keller L, Goodisman MA. 2013. Evolution at two levels in fire ants: the relationship between patterns of gene expression and protein sequence evolution. Mol Biol Evol. 30(2):263–271.
International HapMap 3 Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58.
Jakobsson M, et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451(7181): 998–1003.
Jiang X, Assis R. 2017. Natural selection drives rapid functional evolution of young *Drosophila* duplicate genes. Mol Biol Evol. 34(12):3089–3098.
Jobling M, Hurles M, Tyler-Smith C. 2013. Human evolutionary genetics: origins, peoples & disease. New York (NY): Garland Science.
Jorde L, Watkins WS, Bamshad M. 2001. Population genomics: a bridge from evolutionary history to genetic medicine. Hum Mol Genet. 10(20):2199–2207.
Jurecekova J, et al. 2016. Genome-wide association study of prostate cancer in population of Slovak men. Eur Urol Suppl. 15(11):e1343–e1344.
Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res. 20(10):1313–1326.
Katzmarzyk PT, Leonard WR. 1998. Climatic influences on human body size and proportions: ecological adaptations and secular trends. Am J Phys Anthropol. 106(4):483–503.
Kawashima M, et al. 2017. Genome-wide association studies identify PRKCB as a novel genetic susceptibility locus for primary biliary cholangitis in the Japanese population. Hum Mol Genet. 26(3):650–659.
Keinan A, Mullikin JC, Patterson N, Reich D. 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat Genet. 41(1):66–70.
Kennerdell JR, Carthew RW. 1998. Use of dsRNA-mediated genetic interference to demonstrate that frizzled and frizzled 2 act in the wingless pathway. Cell 95(7):1017–1026.
Kimlin MG. 2008. Geographic location and vitamin D synthesis. Mol Aspects Med. 29(6):453–461.
King M-C, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science 188(4184):107–116.
Kirikoshi H, Sekihara H, Katoh M. 2001. Expression profiles of 10 members of Frizzled gene family in human gastric cancer. Int J Oncol. 19(4):767–771.
Kumar A, Kumar J, Gadodia A, Chumber S, Aggarwal L. 2008. Multiple short-segment colonic duplications. Pediatr Radiol. 38(5):567–570.
Labib K, Tercero JA, Diffley JF. 2000. Uninterrupted MCM2-7 function required for DNA replication fork progression. Science 288(5471):1643–1647.
Lamason RL, et al. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310(5755):1782–1786.
Lappalainen T, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501(7468):506–511.

Leinonen T, Cano J, Mäkinen H, Merilä J. 2006. Contrasting patterns of body shape and neutral genetic divergence in marine and lake populations of threespine sticklebacks. J Evol Biol. 19(6):1803–1812.
Li JZ, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100–1104.
Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30(7):923–930.
Loomis WF. 1967. Skin-pigment regulation of vitamin-D biosynthesis in man: variation in solar ultraviolet at different latitudes may have caused racial differentiation in man. Science 157(3788):501–506.
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15(12):550.
Lutzny G, et al. 2013. Protein kinase c-β-dependent activation of NF-κB in stromal cells is indispensable for the survival of chronic lymphocytic leukemia B cells in vivo. Cancer Cell 23(1):77–92.
MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 42(D1):D986–D992.
Makova KD, Li W-H. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. 13(7):1638–1645.
Martin JF, Bradley A, Olson EN. 1995. The paired-like homeo box gene MHox is required for early events of skeletogenesis in multiple lineages. Genes Del. 9(10):1237–1249.
Mathias RA, et al. 2012. Adaptive evolution of the FADS gene cluster within Africa. PLoS One 7(9):e44926.
McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 5(5):e1000471.
Nei M. 1990. Molecular evolutionary genetics. New York (NY): Columbia University Press.
Nguyen D-Q, Webber C, Ponting CP. 2006. Bias of selection on human copy-number variants. PLoS Genet. 2(2):e20.
Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM. 2004. Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. Mol Biol Evol. 21(7):1308–1317.
Ocaña OH, et al. 2012. Metastatic colonization requires the repression of the epithelial–mesenchymal transition inducer Prrx1. Cancer Cell 22(6):709–724.
Patterson NJ, et al. 2012. Ancient admixture in human history. Genetics 192(3):1065–1093.
Pickrell JK, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 19(5):826–837.
Quiver MH, Lachance J. 2018. Adaptive eQTLs reveal the evolutionary impacts of pleiotropy and tissue-specificity, while contributing to health and disease in human populations. BioRxiv. 444737.
R Core Team. 2013. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Raj A, Rifkin SA, Andersen E, Van Oudenaarden A. 2010. Variability in gene expression underlies incomplete penetrance. Nature 463(7283):913–918.
Ramos PS, Shaftman SR, Ward RC, Langefeld CD. 2014. Genes associated with SLE are targets of recent positive selection. Autoimmune Dis. 2014:203435.
Ranciaro A, et al. 2014. Genetic origins of lactase persistence and the spread of pastoralism in Africa. Am J Hum Genet. 94(4):496–510.
Rees JL. 2003. Genetics of hair and skin color. Annu Rev Genet. 37(1):67–90.
Reyland ME. 2009. Protein kinase C isoforms: multi-functional regulators of cell life and death. Front Biosci (Biosci). 14:2386–2399.
Reynolds J, Weir BS, Cockerham CC. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105(3):767–779.
Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci. 53(1-2):131–147.
Sabeti PC, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832–837.
Sartor MA, et al. 2006. A new method to remove hybridization bias for interspecies comparison of global gene expression profiles uncovers an association between mRNA sequence divergence and differential gene expression in *Xenopus*. Nucleic Acids Res. 34(1):185–200.
Scott GR, Turner CG. 1997. Anthropology of modern human teeth. Cambridge: Cambridge University Press.
Sebat J, et al. 2004. Large-scale copy number polymorphism in the human genome. Science 305(5683):525–528.
Sharp AJ, et al. 2008. A recurrent 15q13. 3 microdeletion syndrome associated with mental retardation and seizures. Nat Genet. 40(3):322–328.
Sheng Y-J, et al. 2011. Follow-up study identifies two novel susceptibility loci PRKCB and 8p11.21 for systemic lupus erythematosus. Rheumatology 50(4):682–688.
Spitze K. 1993. Population structure in *Daphnia obtusa*: quantitative genetic and allozymic variation. Genetics 135(2):367–374.
Stranger BE, et al. 2007. Population genomics of human gene expression. Nat Genet. 39(10):1217–1224.
Sudmant PH, et al. 2015. Global diversity, population stratification, and selection of human copy-number variation. Science 349(6253):aab3761.
Takahashi Y, et al. 2013. Paired related homoeobox 1, a new EMT inducer, is involved in metastasis and poor prognosis in colorectal cancer. Br J Cancer. 109(2):307–311.
Takahata N, Nei M. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics 124(4):967–978.
Takano S, et al. 2016. Prrx1 isoform switching regulates pancreatic cancer invasion and metastatic colonization. Genes Dev. 30(2):233–247.
Troelsen JT, Olsen J, Møller J, SjÖstrÖm H. 2003. An upstream polymorphism associated with lactase persistence has increased enhancer activity. Gastroenterology 125(6):1686–1694.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4(3):e72.
Wallace JA, et al. 2014. Protein kinase C Beta in the tumor microenvironment promotes mammary tumorigenesis. Front Oncol. 4:87.
Wang D, Marsh JL, Ayala FJ. 1996. Evolutionary changes in the expression pattern of a developmentally essential gene in three *Drosophila* species. Proc Natl Acad Sci U S A. 93(14):7103–7107.
Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370.
Weiss LA, et al. 2008. Association between microdeletion and microduplication at 16p11. 2 and autism. N Engl J Med. 358(7):667–675.
Wray GA, et al. 2003. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol. 20(9):1377–1419.
Wright S. 1951. The genetical structure of populations. Ann Eugen. 15(4):323–354.
Yasukochi Y, Satta Y. 2013. Current perspectives on the intensity of natural selection of MHC loci. Immunogenetics 65(6):479–483.
Yi X, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329:75–78.
Yuan Y, Tian L, Lu D, Xu S. 2015. Analysis of genome-wide RNA-sequencing data suggests age of the CEPH/Utah (CEU) lymphoblastoid cell lines systematically biases gene expression profiles. Sci Rep. 5(1):7960.

Zerbino DR, et al. 2018. Ensembl 2018. Nucleic Acids Res. 46(D1):D754–D761.
Zhang H, et al. 2015. Suppression of multidrug resistance by rosiglitazone treatment in human ovarian cancer cells through downregulation of FZD1 and MDR1 genes. Anticancer Drugs 26(7):706–715.
Zhu H, Sun G, Dong J, Fei L. 2017. The role of PRRX1 in the apoptosis of A549 cells induced by cisplatin. Am J Transl Res. 9(2):396–402.

Associate editor: Kirk Lohmueller